ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 98 (2007) 1658 – 1683

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Robust estimation in generalized semiparametric mixed models for longitudinal data

Guoyou Qin[a], Zhongyi Zhu[b,*]

[a]*Biostatistics Department, School of Public Health, Fudan University, Shanghai 200032, PR China*
[b]*Department of Statistics, Fudan University, Shanghai 200433, PR China*

Received 20 November 2005
Available online 20 January 2007

Abstract

In this paper, we consider robust generalized estimating equations for the analysis of semiparametric generalized partial linear mixed models (GPLMMs) for longitudinal data. We approximate the non-parametric function in the GPLMM by a regression spline, and make use of bounded scores and leverage-based weights in the estimating equation to achieve robustness against outliers and influential data points, respectively. Under some regularity conditions, the asymptotic properties of the robust estimators are investigated. To avoid the computational problems involving high-dimensional integrals in our estimators, we adopt a robust Monte Carlo Newton–Raphson (RMCNR) algorithm for fitting GPLMMs. Small simulations are carried out to study the behavior of the robust estimates in the presence of outliers, and these estimates are also compared to their corresponding non-robust estimates. The proposed robust method is illustrated in the analysis of two real data sets.

AMS 1991 subject classification: 62F35; 62G08

Keywords: *B*-spline; Generalized linear models; Longitudinal data; Metropolis algorithm; Mixed model; Newton–Raphson algorithm; Partial linear models; Robustness

1. Introduction

Progress in the research on nonparametric function estimation, generalized linear models [10], mixed models and generalized estimation equations [7] impels the development of semiparametric generalized partial linear mixed models (GPLMMs). In particular, He et al. [4] and Sinha [19]

* Corresponding author. Fax: +8621 65642351.
E-mail address: zhuzy@fudan.edu.cn (Z. Zhu).

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.01.006

studied the robust estimation in generalized partial linear models (GPLM) and generalized linear mixed models (GLMMs), respectively. In this paper, we focus on the robust estimation in GPLMM for longitudinal data.

GPLMM can be viewed as a combination of a GLMM and a fully nonparametric model. GLMMs are popular in the analysis of clustered data including longitudinal data or repeated measurements, and useful for accommodating the overdispersion often observed among non-normally distributed responses and for modeling the dependence among responses inherent in longitudinal or repeated measures data by incorporating random effects [20,25]. It is usually assumed that the random effects have a multivariate normal distribution whose variance components are to be estimated from the data. Moreover, the choice of a partial linear model (PLM) is sometimes to avoid a non-parametric specification of high-dimensional covariates and at other times arises naturally due to categorical covariates (e.g., treatment effects). PLMs are naturally used in such circumstances that the mean or median response is linearly related to some variables but the relation to additional variables are not easily parameterized. Therefore, GPLMM may be helpful to offer further insight into the data with the advantages from GLMM and PLM. A number of authors, including Severini and Staniswalis [18], Härdle et al. [3], and Müller [12], have studied estimation and inference for the GPLM with independent data.

The inference in GPLM or GLMM can be made by maximum likelihood estimation (MLE) and generalized estimating equations (GEE) estimation, but the classical methods originated from MLE and GEE can be sensitive to outliers or departure from underlying distributions. In recent works, He et al. [4] and Sinha [19] discussed the robust estimation of GPLM and GLMM, respectively. Although GPLMMs are widely used in the analysis of clustered data, including longitudinal data or repeated measurements, the study, especially the robust analysis, of GPLMMs has received less attention, possibly due to the increased technical problems imposed by a dependence structure in the data and a partial linear structure in the model.

We consider robust estimation in the framework of GEE. First, we approximate the non-parametric function in GPLMM by B-spline. The spline-based methods have been known to provide optimal rates of convergence for both the parametric and the non-parametric components in a PLM [5]. Under appropriate regularity conditions, we obtain the asymptotic normality of the parametric components and the optimal rate of convergence of the non-parametric components. Second, the GEE approach has some inherent robustness, as it requires no specification of the full likelihood. However, estimating equation such as those considered by Lin and Carroll [9] are highly sensitive to outliers in the sample. Therefore, it is necessary to consider robust estimation. We consider a robust estimating equation similar to He et al. [4], which utilizes a weight function to downweight the effect of leverage points and a bounded score function on the Pearson residuals to limit the influence of outliers in the responses. The proposed work is intended as a robust approach against misspecified likelihood in the sense that the given likelihood may not be true one, e.g. the likelihood will be influenced in the presence of outliers. Finally, the robust estimation in the GPLMMs involves the specification of the posterior distribution of the random effect, which cannot be evaluated in closed form. However, it is possible to approximate the posterior distribution by producing random draws from the distribution using a metropolis algorithm [22], which does not require the specification of the posterior distribution. Here, we adopt the robust Monte Carlo Newton–Raphson (RMCNR) algorithm developed by Sinha [19] to fit GPLMM.

The rest of the paper is organized as follows. In Section 2, we present the model and discuss the approximation of the non-parametric function using the B-spline. Robust estimation under GEE framework is also considered. In Section 3, the asymptotic properties of the robust estimators are investigated, and the consistency and normality of the estimators are obtained. In Section 4, we study two simple binary and poisson partial linear mixed models in simulations. For these two simple models, small simulations based on the stochastic RMCNR and the deterministic robust estimation methods are carried out for investigating the behavior of the robust estimates. The simulation results are reported in Section 4 as well, showing that the robust methods we proposed have good performance. In Section 5, two real data sets are analyzed under binary partial linear mixed models by our robust approach.

2. Models and estimation method

In this paper, we consider a longitudinal study with m subjects and n_i observations over time for the ith subject ($i = 1, \ldots, m$, $j = 1, \ldots, n_i$) for a total of $n = \sum_{i=1}^{m} n_i$ observations. The observed data set is $\{(x_{ij}, y_{ij}, t_{ij}), i = 1, \ldots, m, j = 1, \ldots, n_i\}$. Suppose, conditional on random effect U_i from the ith subject, the elements of observed data vector $Y_i = (y_{i1}, \ldots, y_{in_i})^T$ are drawn from a distribution in the exponential family:

$$f_{Y_i|U_i}(Y_i|U_i, \beta_0, f_0(t), \phi) = \prod_{j=1}^{n_i} \exp[\{y_{ij}\theta_{0,ij} - b(\theta_{0,ij})\}/a(\phi) + c(y_{ij}, \phi)], \tag{2.1}$$

where $\theta_{0,ij}$ are canonical parameters. Let $E(y_{ij}|U_i) = \mu_{0,ij}$, $Var(y_{ij}|U_i) = \phi v(\mu_{0,ij})$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where ϕ is a scale parameter and $v(.)$ is a known variance function. The conditional mean $\mu_{0,ij}$ are related to the canonical parameters through the equation $\mu_{0,ij} = \dot{b}(\theta_{0,ij})$ where $\dot{b}$ denotes the first derivative of b with respect to $\theta_{0,ij}$. For simplicity, we consider $\phi = 1$. The conditional mean $\mu_{0,ij}$ is modeled as

$$\eta_{ij}^0 = g(\mu_{0,ij}) = x_{ij}^T \beta_0 + f_0(t_{ij}) + z_{ij}^T U_i, \quad \mu_{0,ij} = \mu(\eta_{ij}^0) = g^{-1}(\eta_{ij}^0), \tag{2.2}$$

where β_0 is a p-vector regression coefficient with covariates x_{ij}, $f_0(.)$ is an unknown smooth function, the U_i are independent q-vector of random effects associated with covariates z_{ij} and $g(.)$ is a given link function. We assume that the random effect $U = \{U_1, \ldots, U_m\}$ independently follow the same distribution: $U_i \sim f_u(U|\Sigma)$, $i = 1, \ldots, m$, depending on parameter Σ. Furthermore, we assume that the observations from different subjects are independent. Without loss of generality, we also assume t_{ij} are all scaled into the interval [0,1]. In the GPLMMs, the random effects are modeled in the mean $\mu_{0,ij}$, which is different from He et al. [4]. In our models (2.1) and (2.2), there is the non-parametric function f_0, which is the main difference with Sinha [19]. In addition, the special case has been discussed by us that the observed longitudinal data y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ are independent and drawn from a normal distribution conditional on random effect U.

Following He et al. [5], we approximate f_0 by a regression spline. Let $0 = s_0 < s_1 < \cdots < s_{k_n} = 1$ be a partition of the interval [0,1]. Using the s_i as knots, we have $N_k = k_n + l$ normalized B-spline basis functions of order $l + 1$ that form a basis for the linear spline space. Just as pointed out in He et al. [4], regression splines have some desirable properties in approximating a smooth function. It often provides good approximations with a small number of knots. The spline approach also treats a non-parametric function as a linear function with the basis functions as pseudo-design variables, and thus any computational algorithm developed for the GLMMs can be used for the GPLMM.

Let $f_0(t)$ be approximated by $\pi(t)^T \alpha_0$, where $\pi(t) = (B_1(t), \ldots, B_N(t))^T$ is the vector of basis functions, and $\alpha_0 \in R^N$ is the spline coefficient vector. This linearizes our regression model so that our regression problem becomes

$$\eta_{ij}(\theta_0) = g(\mu_{ij}(\theta_0)) = x_{ij}^T\beta_0 + \pi(t_{ij})^T\alpha_0 + z_{ij}^T U_i = D_{ij}^T\theta_0 + z_{ij}^T U_i, \tag{2.3}$$

where $D_{ij} = (x_{ij}^T, \pi_{ij}^T)^T$, and $\theta_0^T = (\beta_0^T, \alpha_0^T)$ is the combined regression parameter vector to be estimated. In matrix notations, we let $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$, $Y_i = (y_{i1}, \ldots, y_{in_i})^T$, where $\mu_{ij} = g^{-1}(D_{ij}^T\theta + z_{ij}^T U_i)$, $D_{ij} = (x_{ij}^T, \pi^T(t_{ij}))^T$, and define X_i, Z_i and π_i in a similar fashion for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. For short, we write v_{ij} instead of $v(\mu_{ij})$ in the following. And β_0, Σ, f_0 are considered to be the true parameters and non-parametric function.

To use only the conditional information on $Y_i|U_i$, we choose a bounded score function ψ and define the following robust estimating equations motivated by the work of He et al. [4] and Sinha [19]:

$$E_{u|y}\left[\sum_{i=1}^{m} D_i^T \Delta_i\{\mu_i(\theta, U_i)\} A_i^{-1/2}\{\mu_i(\theta, U_i)\} h_i\{\mu_i(\theta, U_i)\}\right] = 0, \tag{2.4}$$

where $D_i = (X_i, \pi_i)$ acts as the combined design matrix, $\Delta_i = diag\{\dot{\mu}_{i1}(\theta, U_i), \ldots, \dot{\mu}_{in_i}(\theta, U_i)\}$ with $\dot{\mu}(.)$ denoting the first derivative of $\mu(.)$ evaluated at $D_{ij}^T\theta + z_{ij}^T U_i$, $A_i = A_i\{\mu_i(\theta, U_i)\} = diag\{v_{i1}(\mu_{i1}(\theta, U_i)), \ldots, v_{in_i}(\mu_{in_i}(\theta, U_i))\}$, v_{ij} is the variance function, which is also the function of mean function μ_{ij} in the framework of generalized linear model, $h_i(\theta, U_i) = W_i\{\psi(\mu_i(\theta)) - C_i(\mu_i(\theta))\}$ as the core of the estimating equation with weight matrix W_i and correction terms C_i to be specified later, ψ is considered to be Huber's psi function, $\psi(x) = \min\{c, \max(-c, x)\}$. The tuning constant c is typically chosen to give a certain level of asymptotic efficiency at the underlying distribution. The weighting matrix $W_i = diag\{w_{i1}, \ldots, w_{in_i}\}$ is a diagonal matrix. Similar to Sinha [19], we choose the weight function w_{ij} as a function of the Mahalanobis distance in the form

$$w_{ij} = w(x_{ij}) = \min\left[1, \left\{\frac{b_0}{(x_{ij} - m_x)^T S_x^{-1}(x_{ij} - m_x)}\right\}^{\gamma/2}\right],$$

with $\gamma \geqslant 1$; b_0 is chosen as the 95th percentile of Chi-square distribution with degrees of freedom equal to the dimension of x_{ij}, and m_x and S_x are some robust estimates of location and scale of x_{ij}, such as minimum volume ellipsoid (MVE) estimates of Rousseeuw and van Zomeren [15]. Since $\psi(\mu_i) = \psi(A_i^{-1/2}(Y_i - \mu_i))$, we use $C_i(\mu_i) = E\{\psi(A_i^{-1/2}(Y_i - \mu_i))\}$ to ensure the Fisher consistency of the estimator. Note that the estimating equation (2.4) involves an expectation with respect to the distribution of random effect U conditional on responses Y, which is different from He et al. [4]; whereas the main difference with Sinha [19] is that the dimension of parameter θ to be estimated tends to infinity as $n \to \infty$. In addition, it should be noticed that the expectation, denoted by $E_{u|y}$, in the estimating equation (2.4) includes unknown parameter to be estimated, too. And we will use $E^{(0)}$ to denote the expectation with respect to the true parameters.

Outliers in the response y are usually identified by large residuals, and the Huber's psi function $\psi(x)$ in (2.4) is used to bound the influence of potential outliers in the response. The outliers in the covariates x are generally referred to as "high leverage points", and the weight function w_{ij} in (2.4) is adopted to downweight these design outliers.

The estimates obtained by solving (2.4) are referred to as the robust GEE estimates. Note that the choice of $\psi(r) = r$ and $w(x) = 1$ leads to ordinary GEE estimates.

Following He et al. [4], we use the sample quartiles of $\{t_{ij},\ i=1,\dots,m,\ j=1,\dots,n_i\}$ as knots. For example, if we use three internal knots, they are taken to be the three quartiles of the observed $\{t_{ij}\}$. We use cubic splines (splines of order 4), and the number of internal knots is taken to be the integer part of $N_t^{1/5}$, where N_t is the number of distinct values in $\{t_i,\ i=1,\dots,n\}$. This particular choice is consistent with the asymptotic theory of Section 3, but it is mainly based on our empirical experience and desire for simplicity, and is by no means an optimal choice. The readers are referred to He et al. [4] for details of knots selection of B-spline.

With an initial estimate of θ, we solve (2.4) to find the estimate of θ using the following iterative procedure:

$$\theta^{(i+1)} = \theta^{(i)} + \left\{\sum_{i=1}^{m} D_i^T \Omega_i(\mu_i(\theta, U_i)) D_i\right\}^{-1} \times \left[E_{u|y}\left\{\sum_{i=1}^{m} D_i^T \Delta(\mu_i(\theta, U_i)) A_i^{-1/2}(\mu_i(\theta, U_i)) h_i(\mu_i(\theta, U_i))\right\}\right]\Bigg|_{\theta=\theta^{(i)}}, \tag{2.5}$$

where $\Omega_i(\mu_i(\theta, U)) = -\frac{\partial}{\partial \mu_i}[E_{u|y}\{\Delta(\mu_i(\theta, U_i)) A_i^{-1/2}(\mu_i(\theta, U_i)) h_i(\mu_i(\theta, U_i))\}]\Delta_i(\mu_i(\theta, U_i))$.

Note that, in general, the expectations in (2.5) cannot be computed in closed form as the conditional distribution of $U_i|Y_i$ involves the marginal distribution F_{y_i} of Y_i which cannot be easily computed. The computation is often intractable for complicated problems involving random effects with high dimensions. Here we use an alternative method adopted by Sinha [19] that produces random observations from the conditional distribution of $U_i|Y_i$ by using a Metropolis algorithm (see [22] for details), where the specification of the density f_{y_i} is not required. Then the Monte Carlo approximations to these expectations are used.

In the Metropolis algorithm, we choose f_u as the candidate distribution from which potential new draws are made. Then we specify the acceptance function that provides the probability of accepting the new value (as opposed to retaining the previous value). Let U denote the previous draw from the conditional distribution of $U|Y$, and generate a new value u_j^* for the jth component of $U^* = (u_1, \dots, u_{j-1}, u_j^*, u_{j+1}, \dots, u_{mq})$ by using the candidate distribution f_u. As suggested in McCulloch [11], with probability

$$\alpha_j(U, U^*) = \min\left\{1, \frac{f_{u|y}(U^*|Y, \theta, \Sigma) f_u(U|\Sigma)}{f_{u|y}(U|Y, \theta, \Sigma) f_u(U^*|\Sigma)}\right\} \tag{2.6}$$

accept the candidate value U^*; otherwise, reject it and retain the previous value U. The second term in brace in (2.6) can be simplified to

$$\frac{f_{u|y}(U^*|Y, \beta, \Sigma) f_u(U|\Sigma)}{f_{u|y}(U|Y, \beta, \Sigma) f_u(U^*|\Sigma)} = \frac{f_{y|u}(Y|U^*, \theta)}{f_{y|u}(Y|U, \theta)} = \frac{\prod_{i=1}^m f_{y_i|u_i}(Y_i|U^*, \theta)}{\prod_{i=1}^m f_{y_i|u_i}(Y_i|U, \theta)}.$$

Note that, the calculation of the acceptance function $\alpha_j(U, U^*)$ here involves only the specification of the conditional distribution of $Y|U$ which can be computed in closed form.

Incorporating the Metropolis step into the Newton–Raphson iterative equation (2.5) for the Monte Carlo estimates of the expected values gives an algorithm as follows:

1. Choose initial values $\theta^{(0)}$ and $\Sigma^{(0)}$. These initial estimates can be chosen as the ordinary Monte Carlo Newton–Raphson (MCNR) [11] estimates. Set $m_s = 0$.

2. Generate N observations $U^{(1)}, \ldots, U^{(N)}$ from the distribution $f_{U|Y}(U|Y, \beta^{(m_s)}, \Sigma^{(m_s)})$ using the Metropolis algorithm described previously. Use these observations to find the Monte Carlo estimates of the expectations. Specially,
   (a) Compute $\theta^{(m_s+1)}$from the expression

$$\theta^{(m_s+1)} = \theta^{(m_s)} + \left[\frac{1}{N}\sum_{s=1}^{N}\left\{\sum_{i=1}^{m} D_i^T \Omega_i(\mu_i(\theta^{(m_s)}, U_i^{(s)}))D_i\right\}\right]^{-1}$$
$$\times\left[\frac{1}{N}\sum_{s=1}^{N}\left\{\sum_{i=1}^{m} D_i^T \Delta_i(\theta^{(m_s)}, U^{(s)})A_i^{-1/2}(\mu(\theta^{(m_s)}, U^{(s)}))\right.\right.$$
$$\left.\left.\times\, h_i(\mu_i(\theta^{(m_s)}, U^{(s)}))\right\}\right].$$

   (b) Compute $\Sigma^{(m_s+1)}$ by maximizing

$$\frac{1}{N}\sum_{s=1}^{N} \ln f_u(U^{(s)}|\Sigma).$$

   (c) Set $m_s = m_s + 1$.
3. Continue step 2 until convergence is achieved. Choose $\theta^{(m_s+1)}$ and $\Sigma^{(m_s+1)}$ to be the RMCNR estimates of θ_0 and Σ.

Convergence of this algorithm is not guaranteed, however, the convergence has not been a problem in our empirical investigations for the exact method. As McCulloch [11] pointed out, for sufficiently large simulation sample size, MCNR would inherit the properties of the exact versions.

We explore the behaviors of these Monte Carlo estimates both in the small simulation in Section 4 and in the analysis of two real data sets described in Section 5, and find that the stochastic estimates really provide good approximations to the deterministic ones when the number of replication N is fairly large.

3. Asymptotics properties

Under some regularity conditions, we study the asymptotic properties of $\hat{\beta}$ and $\hat{f}_0$. Meanwhile, If Eq. (2.4) has multiple solutions, only a sequence of consistent estimator $\hat{\theta}$ is considered in this section. A sequence $\hat{\theta}$ is said to be a consistent sequence, if $\hat{\beta}-\beta_0 \to 0$ and $\sup_t |\pi^T(t)\hat{\alpha}-f_0(t)| \to 0$ in probability as $n \to \infty$.

Let $\mu_{0,i}(U_i) = E(Y_i|U_i) = g^{-1}(X_i\beta_0 + f_0(t_i) + Z_iU_i), i = 1, \ldots, m,$ $A_{0,i}(U_i)$ denotes $A_i(\mu_i)$ evaluated at $\mu_i = \mu_{0,i}$, $h_{0,i}(U)$ is in a similar fashion. Also let $e_i = A_{0,i}(U)^{-1/2}(Y_i-\mu_{0,i})$, $h_{0,i}(e_i) = W_i\{\psi(e_i) - E\psi(e_i)\}$. The assumptions required for establishing the asymptotic results similar to those of He et al. [4] and Sinha [19] are as follows:

(A.1) The rth derivative of f_0 is bounded for some $r \geqslant 2$, and suppose $\max_{1\leqslant i\leqslant n} |q_{i+1} - q_i| = o(k_n^{-1})$ and $\max_{1\leqslant i\leqslant n} q_i / \min_{1\leqslant i\leqslant n} q_i \leqslant M$, where $q_i = (s_i - s_{i-1})$, $M > 0$, s_i denotes the ith distinct knot.

(A.2) $\sup_{i\geqslant 1} E^{(0)}\|\mu_{0,i}(U_i) - E^{(0)}\mu_{0,i}(U_i)\|^2 < \infty$.

(A.3) There exists positive constant C_1 such that

$$\infty > v(.) \geqslant C_1 > 0.$$

$g^{-1}(\cdot)$ has bounded third derivatives and $v(.)$ has bounded second derivatives.

To obtain the asymptotic distribution for the estimator $\hat{\beta}$, some assumptions on the covariates X and t are required. One complicating issue for the semiparametric model comes from the dependence between X_i and t_i. To this end, We denote $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ and assume the following relationship as Rice [14]:

$$x_{ijk} = g_k(t_{ij}) + \delta_{ijk}, \quad 1 \leqslant i \leqslant m, \ 1 \leqslant j \leqslant n_i, \ 1 \leqslant k \leqslant p, \tag{3.1}$$

where $g_k(t)$ are functions with bounded rth derivatives, and $\delta'_{ijk}s$ are mean zero random variables independent of $\{e_i\}$, $\{U_i\}$ and of one another. We also assume

(A.4) For sufficiently large n, $k_n(M^T\Omega_0 M)$ is non-singular, and the eigenvalues of $M^T\Omega_0 M$ (k_n/n) are bounded away from zero and infinity, where $M = (\pi_1, \ldots, \pi_n)^T_{n\times N_k}$, $\Omega_0 = diag\{\Omega_{0,i}\}$, $\Omega_{0,i} = -E^{(0)}\{\frac{\partial E_{u|y}\Delta_i(\mu_i)A_i^{-1/2}(\mu_i)h_i(\mu_i)}{\partial \mu_i^T}|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\} = -E^{(0)}[\{E_{u|y}(\frac{\partial \Delta_i(\mu_i)A_i^{-1/2}(\mu_i)h_i(\mu_i)}{\partial \mu_i}) - E_{u|y}(\Delta_i(\mu_i)A_i^{-1/2}(\mu_i)h_i(\mu_i)\mu_i^T A_i^{-1}) + E_{u|y}(\Delta_i(\mu_i)A_i^{-1/2}(\mu_i) h_i(\mu_i)(E_{u|y}A_i^{-1}\mu_i)^T)\}|_{\mu_i=\mu_{0,i}}\Delta_{0,i}]$.

The assumptions (3.1) are first used by Rice [14], then adopted by He et al. [5] and He et al. [4], respectively. Specially, (3.1) ensures achieving the optimal rate of convergence of the estimators $\hat{\beta}$ and $\hat{f}_0$.

(A.5) (a) $E^{(0)}\Lambda_n = 0$ and $\sup_{n\geqslant 1}\frac{1}{n}E^{(0)}\|\Lambda_n\|^2 < \infty$,

(b) $$\frac{1}{n}K_n \xrightarrow{p} K, \quad \frac{1}{n}S_n \xrightarrow{p} S \tag{3.2}$$

for some positive definite matrix K and S.

where Λ_n is n by p matrix, whose sth column is $\delta_s = (\delta_{11s}, \ldots, \delta_{mn_ms})^T$, $S_n = \sum_{i=1}^m X_i^{*T}E^{(0)}\{(E_{u|y}^{(0)}\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i})(E_{u|y}^{(0)}\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i})^T\}X_i^*$, $K_n = \sum_{i=1}^m X_i^{*T}\Omega_{0,i}X_i^*$, $M = (\pi(t_1), \ldots, \pi(t_n))^T$, $X^* = (I-P)X$, $P = M(M^T\Omega_0 M)^{-1}M^T\Omega_0$.

The smoothness condition on f_0 as given by (A.1) determines the rate of convergence of the spline estimate $\hat{f} = \pi(t)^T\hat{\alpha}$ and the distinct values of knots are required to be a quasi-uniform sequence. Higher order derivatives are technically convenient, since they make Taylor expansion possible, but their existence does not seem to be essential for the results to hold. To obtain the asymptotic normality of $\hat{\beta}$, we need (A.5), which are similar to the assumptions by He et al. [4] for achieving the asymptotic properties of robust estimates in GPLM.

It is important to note that the number of distinct knots k has to increase with n for asymptotic consistency. On the other hand, too many knots would increase the variance of our estimators. Therefore, the number of knots must be properly chosen to balance between the bias and variance. For the optimal rate of convergence, we choose $k_n \approx n^{1/(2r+1)}$.

Theorem 1. *Assume conditions* (A.1)–(A.5). *If the number of knots* $k_n \approx n^{1/(2r+1)}$, *then*

$$\frac{1}{n}\sum_{i=1}^m\sum_{j=1}^{n_i}\{\hat{f}(t_{ij}) - f_0(t_{ij})\}^2 = O_p(n^{-2r/(2r+1)}), \tag{3.3}$$

and

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, V_\beta), \tag{3.4}$$

where $V_\beta = K^{-1}SK^{-1}$, *the matrices* K *and* S *are defined in condition* (A.5), $\xrightarrow{\mathcal{L}}$ *denotes convergence in distribution.*

Under rather general conditions (see e.g. [21, Lemmas 8 and 9], (3.3) implies that $\int(\hat{f}(t) - f_0(t))^2\,\mathrm{d}t = O_p(n^{-2r/(2r+1)})$. Under the smoothness condition in (A.1), this is the optimal rate of convergence for estimating f_0. The asymptotic normality (3.4) of $\hat{\beta}$ is useful for making large sample inference on β_0. To do so, we present the estimate of the asymptotic covariance matrix of the robust GEE as follows:

$$\hat{V}_\beta = n\hat{K}_n^{-1}\hat{S}_n\hat{K}_n^{-1}, \tag{3.5}$$

where $\hat{K}_n$ and $\hat{S}_n$ are obtained as

$$\hat{K}_n = \sum_{i=1}^{m} X_i^{*T}\Omega_i X_i^*, \tag{3.6}$$

and

$$\hat{S}_n = \sum_{i=1}^{m} X_i^{*T}\{E_{u|y}(\Delta_i A_i^{-1/2}h_i)\}\{E_{u|y}(\Delta_i A_i^{-1/2}h_i)\}^T X_i^*, \tag{3.7}$$

where $h_i = (h_{i1}(\theta), \ldots, h_{in_i}(\theta))^T$, and all the quantities involved are evaluated at $\hat{\theta}$.

By the following Theorem 2, the asymptotic covariance matrix $V_\beta = K^{-1}SK^{-1}$ can be consistently estimated by $\hat{V}_\beta$.

Theorem 2. *Under the conditions of Theorem* 1, *if the number of knots* $k_n \approx n^{1/(2r+1)}$, *then*

$$n^{-1}\hat{K}_n \xrightarrow{p} K, \quad n^{-1}\hat{S}_n \xrightarrow{p} S. \tag{3.8}$$

Theorems 1 and 2 are established in the case of no outliers, however, in the presence of outliers, they may not hold any more. As indicated by Huber [6], a robust procedure is expected to have good efficiency at the assumed model with no outliers and the insensitivity to small deviation from the model assumptions. From our simulations, it could be found that the proposed robust estimator possesses such desirable features.

4. Simulation study

To evaluate the performance of robust GEE method, two sets of small simulation studies, respectively, fitting simple binary and Poisson partial linear mixed models are conducted. Note that as there is only simple random effect in the following binary and Poisson partial linear mixed models (4.1) and (4.4) which are specified later, it is relatively easy to find the exact robust GEE estimates by evaluating the integrals involving the conditional expectations using numerical methods but not Monte Carlo estimates. Therefore, in the simulations, we consider finding both

Table 1
Simulation results for binary response in Study 1 over 500 replications

		IMSE	BIAS($\hat{\beta}$)	MCse($\hat{\beta}$)	$\sqrt{MSE(\hat{\beta})}$	AEse($\hat{\beta}$)	BIAS($\hat{\sigma}^2$)	MCse($\hat{\sigma}^2$)	AEse($\hat{\sigma}^2$)
NP	NR	0.1539	0.0463	0.2611	0.2651	0.2451 (0.0220)	0.0931	0.5080	0.4425
	R	0.1634	0.0458	0.2699	0.2737	0.2487 (0.0230)	0.0920	0.5147	0.4420
P	NR	0.1486	−0.3157	0.3381	0.4626	0.3410 (0.0675)	−0.0427	0.4670	0.4168
	R	0.1842	0.0739	0.3060	0.3148	0.2553 (0.0255)	0.1950	0.5828	0.4598

NP, no perturbation; P, with perturbation; NR, non-robust method; R, robust method; MCse, Monte Carlo standard error; AEse, average estimated standard error.

exact robust GEE and RMCNR estimates. Here, Simpson integration method is used to find the exact robust GEE estimate.

The bias, standard error and the square root of MSE of the robust GEE estimates $\hat{\beta}$ as well as the integrated mean squared error (IMSE) of $\hat{f}$ will be estimated and compared with their corresponding non-robust GEE ones defined through the same estimating equations except that $w_i = I$ and $\psi(x) = x$ both in the absence and in the presence of outliers.

In our simulation, the γ in the weight function is chosen to be 1 and the tuning constant c of Huber's psi function is chosen to be 1.5 as suggested by He et al. [4].

Study 1: We consider a binary partial linear mixed model with a single random effect, a single fixed effect and a single non-parametric function:

$$y_{ij}|U \sim independent\ Bernoulli(\mu_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

$$\eta_{ij} = \log \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_1 x_{ij} + 1.5\,\cos(t_{ij}\pi/2) + u_i, \quad u_i \sim N(0, \sigma^2), \tag{4.1}$$

where $m = 100$, $n_i = 4$, $\beta_1 = 1.5$, $\sigma^2 = 1$, and x_{ij} are drawn independently from uniform distribution on $(-1, 1)$, whereas t_{ij} are drawn from uniform distribution on $(0, 1)$ independent of x_{ij}. A total of 500 samples are drawn from model (4.1).

To study robustness, similarly to Sinha [19], some outliers are created in the data set by moving four randomly chosen points (i.e., 1%) from the bulk of the data toward x direction. More specially, to create the outliers, we replace the corresponding x value by $x + 5$. These type of outliers are referred to as mean shift outliers, and the ordinary estimates are often heavily influenced by such outliers. Table 1 compares the performance of the robust estimators against non-robust ones by exact method. The number of internal knots is taken to be 3, the integer part of $400^{1/5}$. In Table 1, it is observed that both the non-robust and the robust estimators perform almost equally well in the case of no outliers, although we lose some efficiency in the robust method with a slightly larger biases and MSEs of the parameter estimates. This is a small premium one needs to pay for using the robust method when there is, in fact, no outliers in the data. However, the main purpose of this study is to explore the performance of the proposed robust method in the presence of outliers. Table 1 also presents that the outliers do not appear to have any serious impact on the biases and MSEs of the robust estimates in the case of outliers, in the contrast, the non-robust estimates are generally seen to be heavily affected by the outliers as the corresponding biases and MSEs are large in magnitude. We also observe that the non-robust estimates of the variance component σ^2 appear to have smaller biases and MSEs than the robust ones, which is deserved further study.

Table 2
Simulation results for binary response in Study 1 over 500 replications by RMCNR and MCNR

		IMSE	BIAS($\hat{\beta}$)	MCse($\hat{\beta}$)	$\sqrt{MSE(\hat{\beta})}$	AEse($\hat{\beta}$)	BIAS($\hat{\sigma}^2$)	MCse($\hat{\sigma}^2$)	AEse($\hat{\sigma}^2$)	RBIAS($\hat{\sigma}^2$)	RMCse($\hat{\sigma}^2$)
NP	NR	0.1536	0.0472	0.2607	0.2650	0.2445 (0.0231)	0.0998	0.5087	0.4489	0.1037	0.5139
	R	0.1623	0.0439	0.2707	0.2742	0.2479 (0.0246)	0.0878	0.5369	0.4432	0.0910	0.5407
P	NR	0.1482	−0.3150	0.3390	0.4627	0.3372 (0.0681)	−0.0348	0.4637	0.4198	−0.0328	0.4665
	R	0.1825	0.0678	0.3059	0.3132	0.2504 (0.0260)	0.1797	0.6268	0.4600	0.1824	0.6319

NP, no perturbation; P, with perturbation; NR, non-robust method; R, robust method; MCse, Monte Carlo standard error; AEse, average estimated standard error.

Table 3
Simulation results for Poisson response in Study 2 over 500 replications

		IMSE	BIAS($\hat{\beta}$)	MCse($\hat{\beta}$)	$\sqrt{MSE(\hat{\beta})}$	AEse($\hat{\beta}$)	BIAS($\hat{\sigma}^2$)	MCse($\hat{\sigma}^2$)	AEse($\hat{\sigma}^2$)
NP	NR	0.0130	0.0089	0.1256	0.1261	0.1240 (0.0123)	−0.0021	0.0555	0.0518
	R	0.0135	0.0082	0.1282	0.1285	0.1269 (0.0113)	−0.0024	0.0555	0.0517
P	NR	0.0141	−0.3515	0.1852	0.3974	0.2022 (0.0437)	0.0007	0.0560	0.0521
	R	0.0146	−0.0266	0.1321	0.1349	0.1254 (0.0108)	0.0059	0.0581	0.0523

NP, no perturbation; P, with perturbation; NR, non-robust method; R, robust method; MCse, = Monte Carlo standard error; AEse, average estimated standard error.

We also computed the standard errors of the parameter estimates using the large-sample approximation (3.5). The expectations in (3.6) and (3.7) can be approximated by their Monte Carlo estimates. The variance of the estimates of σ^2 is also computed from the observed Fisher information for σ^2. It is observed that the standard error of the proposed robust estimator is underestimated by the asymptotic approximation given by Theorem 2 for the robust method especially in the presence of outliers. In general, the sandwich estimator for the covariance by GEE method usually results in the underestimated standard error. However, it is slightly serious in the presence of outliers, which is a topic and need further investigation. The average estimated standard errors of $\hat{\beta}$ ($AEse(\hat{\beta})$) and $\hat{\sigma}^2(AEse(\hat{\sigma}^2))$ in Table 1 are compared to the empirical standard error of $\hat{\beta}(MCse(\hat{\beta}))$ and $\hat{\sigma}^2$ ($MCse(\hat{\sigma}^2)$) based on the 500 samples, respectively. The numbers in parentheses are the empirical standard error of $AEse(\hat{\beta})$. We also note that, in the presence of outliers, the large sample standard error estimates are much more stable for the robust estimators.

Table 2 presents the simulation results by RMCNR and MCNR from which we can get similar conclusions as in Table 1. In McCulloch [11], the replication N, which is called Monte Carlo sample size, is increased with the number of iteration through the *ad hoc* method and a predetermined number of iterations is used. Just as pointed out in McCulloch [11], the Monte Carlo estimates reach the neighborhood of the exact estimates quickly, but they continue to show random variation. And the number of replications N required to get stochastic estimates to converge with four or three-decimal accuracy should be very large, which will result in time consuming. Therefore, for simplicity and time saving, the Monte Carlo sample size N is chosen to be 500 and the number of iterations is predetermined to be 30, which results in about two-decimal accuracy in the simulation study. From Table 2, it is found that the stochastic estimates provide good approximations to the deterministic ones.

Note that the estimate of σ^2 (step 2(b) of the Metropolis algorithm) can be updated as

$$\sigma^{2(m_s+1)} = \frac{1}{N}\sum_{s=1}^{N}\frac{1}{m}U^{(s)T}U^{(s)}.$$

And to study the robustness of the estimate of variance component σ^2 proposed here, we attempt to compare it with the robust one through the following median absolute deviation (MAD)

$$\hat{\sigma}_r^2 = [1.4826 * median\{|U - median(U)|\}]^2,$$

where U is drawn from the posterior distribution of $U|Y$. The bias (RBIAS(σ^2)) and empirical standard error (RMcse(σ^2)) of such estimate are also given in Table 2, which are close to those of our estimates. It seems that the estimate of σ^2 proposed here is robust in some extent.

Table 4
Simulation results for Poisson response in Study 2 with design outliers over 500 replications by RMCNR and MCNR

		IMSE	BIAS($\hat{\beta}$)	MCse($\hat{\beta}$)	$\sqrt{MSE(\hat{\beta})}$	AEse($\hat{\beta}$)	BIAS($\hat{\sigma}^2$)	MCse($\hat{\sigma}^2$)	AEse($\hat{\sigma}^2$)	RBIAS($\hat{\sigma}^2$)	RMCse($\hat{\sigma}^2$)
NP	NR	0.0138	0.0091	0.1272	0.1277	0.1269 (0.0199)	−0.0155	0.0523	0.0504	−0.0102	0.0572
	R	0.0139	0.0080	0.1292	0.1292	0.1272 (0.0122)	−0.0152	0.0526	0.0504	−0.0098	0.0574
P	NR	0.0159	−0.3601	0.1869	0.4046	0.2042 (0.0477)	−0.0128	0.0516	0.0508	−0.0072	0.0556
	R	0.0147	−0.0297	0.1322	0.1428	0.1271 (0.0186)	−0.0087	0.0538	0.0510	−0.0026	0.0581

NP, no perturbation; P, with perturbation; NR, non-robust method; R, robust method; MCse, Monte Carlo standard error; AEse, average estimated standard error.

Table 5
Simulation results for Poisson response in Study 2 with response outliers over 500 replications by Exact and MCNR methods

		IMSE	BIAS($\hat{\beta}$)	MCse($\hat{\beta}$)	$\sqrt{MSE(\hat{\beta})}$	AEse($\hat{\beta}$)	BIAS($\hat{\sigma}^2$)	MCse($\hat{\sigma}^2$)	AEse($\hat{\sigma}^2$)	RBIAS($\hat{\sigma}^2$)	RMCse($\hat{\sigma}^2$)
Exact	NR	0.0242	−0.0803	0.1567	0.1761	0.1550 (0.0200)	0.0319	0.0560	0.0564	–	–
	R	0.0280	−0.0098	0.1422	0.1425	0.1333 (0.0133)	0.0850	0.0713	0.0620	–	–
MCNR	NR	0.0277	−0.0792	0.1566	0.1755	0.1630 (0.1030)	0.0200	0.0540	0.0552	0.0328	0.0626
	R	0.0223	−0.0095	0.1417	0.1421	0.1347 (0.0163)	0.0602	0.0638	0.0593	0.0678	0.0705

Exact, exact method; MCNR, Monte Carlo Newton–Raphson method; NP, no perturbation; P, with perturbation; NR, non-robust method; R, robust method; MCse, Monte Carlo.

Study 2: This is a similar set-up as Study 1, but (4.1) is replaced by

$$y_{ij}|U \sim independent P(\lambda_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$
$$\eta_{ij} = \log(\lambda_{ij}) = \beta_1 x_{1,ij} + \sin(2t_{ij}) + u_i, \quad u_i \sim N(0, \sigma^2), \tag{4.2}$$

where $\beta_1 = 1$, $\sigma^2 = 0.25$, and x_{ij} are drawn independently from uniform distribution on $(-0.5, 0.5)$, whereas t_{ij} are drawn from uniform distribution on (0, 1) independent of x_{ij}. The outliers created here are similar to Study 1 expect the corresponding x_{ij} value are replaced by $x_{ij} - 3$. In Tables 3 and 4, we obtain the similar conclusions as in Study 1.

A referee pointed out that in the above simulations, only design outliers are investigated. It would be more interesting to see how the robust estimates behave in the presence of response outliers. Here, we carry out a separate simulation to investigate the behavior of the proposed robust estimates when the data are contaminated with response outliers in the semiparametric Poisson mixed model in Study 2. Note that the outliers can arise only through the x values in the binary mixed model in Study 1 as the response y is binary. In the following simulation, the set-up is the same as that in Study 2. But the outliers are created by replacing 8 randomly chosen y_{ij} values by $y_{ij} + 10$. Table 5 presents the simulation results over 500 replications by both the exact and MCNR methods. As expected, the proposed robust estimates provide smaller bias and MSE than the non-robust ones in the presence of response outliers. However, the non-robust estimates of the variance component σ^2 appear to have smaller biases and MSEs than the robust ones, which is similar to that occurs in Study 1. And the stochastic estimates also provide good approximation to the deterministic ones. The conclusions are similar to those in Study 1.

5. Examples

To further illustrate the effectiveness of the proposed method in this paper, we apply the GPLMM and the robust estimating equations (2.4) to two real data sets. As pointed out by McCulloch [11], the Monte Carlo sample size N required to get stochastic estimates to converge with four or three-decimal accuracy should be very large. Therefore, in the analysis of each of the two real data sets, the Monte Carlo sample size N is chosen to be 2000 and the number of iterations is predetermined to be 100 for higher accuracy.

Example 1 (*GUIDE study of Preisser and Qaqish [13]*). Preisser and Qaqish [13] analyzed an interesting set of data from Guidelines for Urinary Incontinence Discussion and Evaluation. A total of 137 patients of age 76 or above who had experienced accidential loss of urine and had been using some of the 38 medical practices were asked whether they were bothered by the problem. The binary response variable y_{ij} is 1 if the ith patient from the jth medical practice is "bothered" by the urinary incontinence, and 0 otherwise. A conditional independent logistic model was used by Sinha [19] with the following five covariates: standardized age (AGE), GENDER (1 = female), the number of leaking accidents per day (DAYACC), severity of leaking (SEVERE) on a scale of 1–4 (1 = just create some moisture, 2 = wet their underwear, 3 = trickle down their thigh, and 4 = wet the floor), and the number of times during the day they usually go to toilet to urinate (TOILET). The standardized age is (age (in years) $-$ 76)/10.

We fit the data by the following binary partial linear mixed model:

$$logit(\mu_{ij}) = X_{ij}^T\beta + f_0(\text{AGE}) + u_i, \tag{5.1}$$

where $u_i \overset{iid}{\sim} N(0, \sigma^2)$.

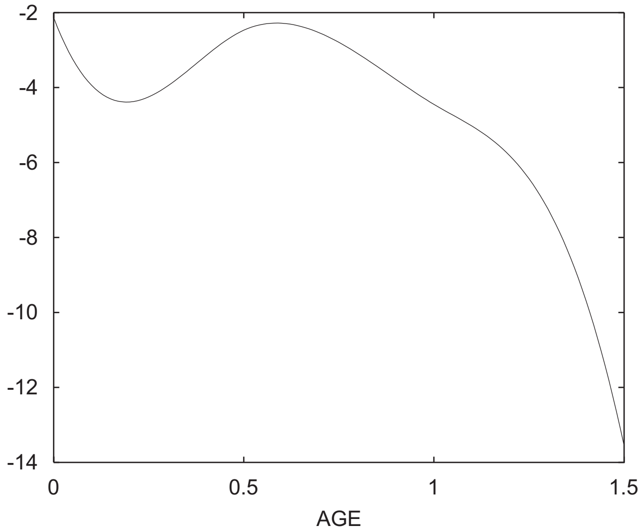

Fig. 1. The estimated function on AGE for the GUIDE study.

Table 6
Regression coefficient estimates in analysis of the GUIDE data

	Semiparametric model				Parametric model
	Robust	RMCNR	Non-robust	H,F&Z	Sinha
Intercept	–	–	–	–	−3.5928 (0.9519)
Gender	−1.5942 (0.6026)	−1.5901 (0.6081)	−1.1590 (0.6577)	−1.57(0.61)	−1.2977 (0.6315)
Age (10 years)	–	–	–	–	−1.0724 (0.6234)
Dayacc	0.6680 (0.1488)	0.6676 (0.1496)	0.6151 (0.1360)	0.59 (0.14)	0.5061 (0.1161)
Severe	0.8264 (0.4643)	0.8199 (0.4669)	1.0939 (0.4507)	0.67 (0.40)	0.8274 (0.3728)
Toilet	0.1315 (0.0994)	0.1324 (0.1002)	0.0819 (0.0894)	0.27 (0.10)	0.2396 (0.1101)
Variance	1.9841 (1.3526)	1.9569 (1.3085)	1.7022 (1.2311)	–	1.8605 (1.4136)
Robust variance	–	1.9179	–	–	–

The model applied here is similar to Sinha [19] except the AGE variable entering the model as a four order regression spline with two internal knot, and the random effect u_i in (5.1) is the only difference with He et al. [4].

The estimated function on AGE is given in Fig. 1, which indicates an interesting non-liearity: After 85 years old, the probability of being bothered by the accidential loss of urine decreases with age, which is similar to He et al. [4].

Table 6 gives the results of our study in comparison with the estimates of He et al. [4] and Sinha [19]. Due to the difference in how AGE is included in the model, the weight functions in our robust method are computed from $Z =$ (DAYACC, TOILET). In Table 6, it is found that the GENDER and DAYACC effects are both significant at the 0.05 level by both the robust method and non-robust one. However, the SEVERE effect is insignificant by the robust method. The p value of the test $\mathrm{H}_0 : \sigma^2 = 0$ vs $\mathrm{H}_1 : \sigma^2 > 0$ indicates that the variance component σ^2 is

Table 7
Regression coefficient estimates in analysis of the infectious disease data

	GPLMM			GPLM	
	Robust	RMCNR	Non-robust	H,F&Z	L&C
Vitamin A	0.4476 (0.4373)	0.4515 (0.4405)	0.5001 (0.4486)	0.715 (0.367)	0.611 (0.529)
Seasonal cosine	−0.7172 (0.1893)	−0.7127 (0.1898)	−0.6130 (0.1833)	−0.698 (0.188)	−0.587 (0.210)
Seasonal sine	−0.1202 (0.1684)	−0.1207 (0.1686)	−0.1621 (0.1541)	−0.079 (0.166)	−0.161 (0.183)
Sex	−0.6827 (0.2816)	−0.6810 (0.2834)	−0.5536 (0.2570)	−0.611 (0.269)	−0.508 (0.295)
Height	−0.0465 (0.0390)	−0.0458 (0.0393)	−0.0295 (0.0279)	−0.069 (0.039)	−0.026 (0.035)
Stunting	0.3881 (0.5005)	0.3841 (0.5048)	0.4976 (0.4307)	0.071 (0.466)	0.463 (0.525)
Variance	0.6210 (0.2623)	0.5570 (0.2364)	0.5550 (0.2525)	–	–
Robust variance	–	0.5575	–	–	–

significant at 0.0712 level by the robust method, whereas it is significant at 0.0834 level by the non-robust one.

As pointed by Sinha [19], the potential influential observations may include the patients 7, 10, 27, 56, 59, 97 and 131. Particularly, the patient 97 appears to be the most extreme point with smallest weight. Here, to get some idea about the potential influential observations in the data, following Sinha [19], we calculate the weight function used in our robust method as $s_{ij} = E_{u|y}[\{\psi_c(r_{ij}(\hat{\theta}, U_i)) - E_{y|u}\{\psi_c(r_{ij}(\hat{\theta}, U_i))\}\}w_{ij}\dot{\mu}_{ij}(\hat{\theta}, U_i))/v_{ij}^{1/2}(\hat{\theta}, U_i)]/E_{u|y}\{r_{ij}(\hat{\theta}, U_i)\dot{\mu}_{ij}(\hat{\theta}, U_i)/v_{ij}^{1/2}(\hat{\theta}, U_i)\}$, where $r_{ij}(\hat{\theta}, U_i) = (y_{ij} - \mu_{ij}(\hat{\theta}, U_i))/v_{ij}^{1/2}(\hat{\theta}, U_i)$. Note that for the choice $w_{ij} = 1$ and $\psi_c(r_{ij}) = r_{ij}$, the weight function $s_{ij} = 1$ and the robust estimation reduces to the non-robust one. The heavily downweighted points (with weights less than 0.10) include the patients 10, 45, 47, 56, 59, 97, 98 and 131. Patients 97 reports SEVERE = 3, DAYACC = 16.7, and TOILET = 8 and appeared to be the most extreme point with the smallest weight $s_{ij} = 0.0053$. The result of the analysis of the potential influential observations is consistent with that in Sinha [19]. The robust method downweights those subjects and more accurately reflects the relationship in the majority of patients.

We also computed the RMCNR estimates, which are the Monte Carlo version of the exact robust GEE estimates. Following Sinha [19], we use $N = 2000$ replicates in the iterative equation of the RMCNR method so that the estimates can be compared with two-decimal accuracy. These results are shown in Table 6 as well. As expected, the stochastic RMCNR estimates appear to be very close to the deterministic exact robust GEE estimates.

Example 2 (*An infectious disease study*). Similar to Example 1, a partial linear logistic-normal random effect model is fitted to infectious disease data on 275 Indonesian children. The preschool children were examined every 3 or 18 months for the presence of respiratory infection. The response variable is the presence of respiratory infection ($1 = \text{yes}, 0 = \text{no}$), and the covariates of interest include: Vitamin A deficiency ($1 = \text{yes}, 0 = \text{no}$), age, sex ($1 = \text{female}, 0 = \text{male}$), height for age, stunting status (1=yes, 0=no), and seasonable cosine and seasonable sine variables.

He et al. [4] applied GPLM to this data with age entering the model non-parametrically. Here, we also use a four order regression spline with two internal knots to approximate the non-parametric function. Table 7 compares the estimates under GPLMM with those under GPLM by He et al. [4] and Lin and Carroll [9]. Only height for age is continuous variable used in computing the weights in our robust equation. In Table 7, seasonal cosine and sex are the only two significant effects while the other four effects are insignificant at level 0.05, which is the same as Lin and Carroll [9].

The parameter estimates and the SEs are quite similar between robust and non-robust methods which is not surprising when the data contained no outliers. And the p value of the test H_0 : $\sigma^2 = 0$ vs $\mathrm{H}_1 : \sigma^2 > 0$ indicates that the variance component σ^2 is highly significant by both the robust method and the non-robust one, which is the same as Zeger and Karim [24]. The estimated function of AGE using our robust estimating equation also looks like to that of Lin and Carroll [8, Figure 3], so we omit the figure here.

6. Discussion

This paper considers the robust estimating equation of GPLMM for longitudinal data. Bounded score functions and leverage-based weights are used in the estimating equation to achieve robustness against outliers and influential data points. In practical implementation, Monte Carlo Newton–Raphson (MCNR) algorithm is used to approximate intractable integrals due to the conditional distribution of random effects given observed data. The GPLMM considered in this paper extends the model GPLM in He et al. [4] by incorporating the random effects to model the dependency within the subject observations for longitudinal data, which results in a more complicate estimating equation with the mathematical expectation. Sinha [19] studies robust estimation in GLMM and the GPLMM can also be looked as an extension of GLMM by incorporating a non-parametric function used to describe the non-linear relationship between the response and covariates. In practice, semiparametric models are used widely in the data analysis, which also increase the difficulty in study as the dimension of the parameters to be estimated by estimating equations will tend to infinity as $n \to \infty$.

Theorems 1 and 2, which present the asymptotic properties of the proposed robust estimator, are established in the case of no outliers although they may not hold in the presence of outliers. The influence function of the proposed robust estimator is bounded because bounded score function are used. Furthermore, leverage-based weights are adopted to limit the influence of high leverage points in the covariates. Therefore, the proposed robust estimator would be insensitive to small deviations from the assumed model and useful to deal with the outliers, which are also demonstrated by the simulation study. The asymptotic properties established at the assumed model with no outliers and the insensitivity to small deviation from the model assumptions are some desirable features which a robust procedure should achieve [6].

The expectation C_i in the estimating equation used to ensure the Fisher consistency are not available in general unless the true likelihood function is known. In the generalized mixed model setting, C_i can be calculated easily for binary data as y_{ij} only take value 0 or 1 while the C_i are difficult to obtain for the data following other distributions as the calculation of expectation C_i involves intractable integrals. Some numerical integration methods or approximation ones are required to achieve the expectation C_i in this situation. In practice, if the information about the distribution cannot be obtained by experience or the mechanism of the generation of data, an alternative method proposed by Wang et al. [23] can be used, which provides a bias correction method for robust estimation functions without need of assumption of the distribution of the data.

The choice of Monte Carlo sample size N and how to set up stopping rules are important and need further investigation. Because the simulation is time consuming, there is an obvious trade-off between accurate approximation and speed. As noted earlier in Section 4, the *ad hoc* method is used for increasing the Monte Carlo size N and the number of iterations is predetermined for the Markov Chain Monte Carlo algorithm in McCulloch [11]. Furthermore, Booth and Hobert

[1] discussed the automated procedures to improve the estimation of GLMM upon deterministic choices of Monte Carlo sample size N as well as the stopping rules, in which an appropriate value for N is chosen after each iteration and the algorithm is stopped when changes in the parameter estimates are small after taking Monte Carlo error into account. Their method can be more efficient than that based on Markov Chain Monte Carlo algorithm except that the intractable integrals in the likelihood function are of high dimensions. How to develop a similar automated procedure for the methods based on Markov Chain Monte Carlo algorithm is a challenge and deserves careful investigation.

Acknowledgment

The authors are grateful to the Editor and two referees for their constructive suggestions that largely improve the presentation of the paper and thank the support of the Natural Science Foundation of China (10371042, 10671038).

Appendix A. Proof

To prove our main results, some preliminary lemmas are needed.

Lemma A.1. *Under condition* (A.1), *there exists a constant C_3 depending only on l, and C_0 such that*

$$\sup_{t\in[0,1]} |f_0(t) - \pi^T(t)\alpha_0| \leqslant C_3 k_n^{-r},$$

where α_0 is a N-dimensional vector depending on f_0.

The proof of this lemma follows readily from Schumaker [16, Theorem 12.7].

By Lemma A.1, we approximate $f_0(t)$ by $\pi^T(t)\alpha_0$, then have

$$\eta_{ij}(\theta_0) = g(\mu_{ij}(\theta_0)) = x_{ij}^T\beta_0 + \pi_{ij}^T(t_{ij})\alpha_0 + z_{ij}^T U_i, \quad \theta_0 = (\beta_0^T, \alpha_0^T)^T_{(p+N_k)\times 1}.$$

Proof of Theorem 1. Similar to He et al. [4], let

$$\xi(\beta, \alpha) = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} K_n^{1/2}(\beta - \beta_0) \\ k_n^{-1/2} H_n(\alpha - \alpha_0) + k_n^{1/2} H_n^{-1} M^T \Omega_0 X(\beta - \beta_0) \end{bmatrix},$$

$$\widehat{\xi} = \xi(\widehat{\beta}, \widehat{\alpha}) = \begin{bmatrix} \widehat{\xi}_1 \\ \widehat{\xi}_2 \end{bmatrix},$$

where $H_n^2 = k_n M^T \Omega_0 M$. We shall show that $\|\widehat{\xi}\| = O_p(k_n^{1/2})$. To do so, we standardize $\widetilde{X}_i^T = K_n^{-1/2} X_i^{*T}$, $\widetilde{\pi}_i^T = k_n^{1/2} H_n^{-1}\pi(t_i)^T$, $R_{ni} = \pi(t_i)\alpha_0 - f_0(t_i)$ and $\zeta_i = \widetilde{X}_i\xi_1 + \widetilde{\pi}_i\xi_2 + R_{ni}$, then $g(\mu_i(\theta)) = D_i\theta + Z_i U_i = \eta_{0,i}(U_i) + \zeta_i$, $i = 1, \ldots, m$, where $\eta_{0,i}(U_i) = X_i\beta_0 + f_0(t_i) + Z_i U_i$. Then estimating equation (2.4) become

$$U_\xi(\mu(\xi)) = E_{u|y}[U_\xi(\mu(\xi, U)] = E_{u|y}\left[\sum_{i=1}^m D_i^T \Delta_i A_i^{-1/2} h_i(\mu_i(\xi, U_i))\right] = 0. \tag{A.1}$$

Let

$$T=\begin{bmatrix} K_n^{-1/2} & -K_n^{-1/2}X^T\Omega_0M(M^T\Omega_0M)^{-1} \\ 0 & k_n^{1/2}H_n^{-1}\end{bmatrix},$$

then the robust estimating equation can be written as

$$\begin{aligned}\Psi(\mu(\xi)) &= \begin{bmatrix}\Psi_1(\mu(\xi))\\ \Psi_2(\mu(\xi))\end{bmatrix} = E_{u|y}[TU_\xi(\mu(\xi,U))] \\ &= \begin{bmatrix} E_{u|y}\left\{\sum_{i=1}^m K_n^{-1/2}X_i^{*T}\Delta_iA_i^{-1/2}h_i(\mu_i(\xi,U))\right\} \\ E_{u|y}\left\{\sum_{i=1}^n k_n^{-1/2}H_n^{-1}\pi_i^T\Delta_iA_i^{-1/2}h_i(\mu_i(\xi,U))\right\}\end{bmatrix} \\ &= \sum_{i=1}^n E_{u|y}\{\widetilde{D_i}^T\Delta_iA_i^{-1/2}h_i(\mu_i(\xi,U))\}=0, \end{aligned} \tag{A.2}$$

where $\widetilde{D_i}=(X_i^*K_n^{-1/2},\pi_iH_n^{-1}k_n^{1/2})$. (A.4) and (A.5) guarantee that both (6.1) and (6.2) give the same root for ξ as the estimate. Furthermore, we write

$$\Phi(\xi)=\begin{bmatrix}\Phi_1(\mu(\xi))\\ \Phi_2(\mu(\xi))\end{bmatrix}=\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}+E_{u|y}^{(0)}\left\{\sum_{i=1}^m \widetilde{D}_i^T\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i}(e_i)\right\}. \tag{A.3}$$

The zero $\widetilde{\xi}$ of $\Phi(\xi)$

$$\widetilde{\xi}=\begin{bmatrix}\widetilde{\xi}_1\\ \widetilde{\xi}_2\end{bmatrix}=-\begin{bmatrix} E_{u|y}^{(0)}\left\{\sum_{i=1}^m K_n^{-1/2}X_i^{*T}\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i}(e_i)\right\} \\ E_{u|y}^{(0)}\left\{\sum_{i=1}^m k_n^{1/2}H_n^{-1}\pi_i^T\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i}(e_i)\right\}\end{bmatrix} \tag{A.4}$$

is not a estimate, but we shall prove the difference between $\widehat{\xi}$ and $\widetilde{\xi}$ is small. To do so, let $a\in R^{p+N_k}$ satisfying $a^Ta=1$. We expand $a^T\Psi(\xi)$ in a Taylor series

$$\begin{aligned} a^T\Psi(\mu(\xi)) &= a^T\Psi(\mu(\eta_0+\zeta)) \\ &= E_{u|y}\left\{\sum_{i=1}^m a^T\widetilde{D_i}^T\Delta_iA_i^{-1/2}h_i(\mu_i(\eta_{0,i}+\zeta_i))\right\} \\ &= \sum_{i=1}^m a^T\widetilde{D_i}^TE_{u|y}^{(0)}\Delta_{0,i}A_{0,i}^{-1/2}h_{0,i}(e_i(U)) \\ &\quad+\sum_{i=1}^m a^T\widetilde{D_i}^T\left.\frac{\partial E_{u|y}\Delta_iA_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\zeta_i+R_n^*(\mu^*), \end{aligned}$$

where $R_n^*(\mu)=\sum_{i=1}^m R_{n,i}^*(\mu_i^*)$, $R_{n,i}^*(\mu_i)=\frac{1}{2}\zeta_i^T\Delta_i^T\frac{\partial^2a^T\widetilde{D_i}^TE_{u|y}\Delta_iA_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^2}\Delta_i\zeta_i$ evaluated at $\mu_i^*=g^{-1}(\eta_{0,i}+\tau_i\zeta_i)$, $i=1,\ldots,m$, $0<\tau_i<1$. Then the difference between $a^T\Psi(\mu(\xi))$ and $a^T\Phi(\xi)$ can be expressed as

$$a^T(\Psi(\xi)-\Phi(\xi))=\sum_{i=1}^m\left[a^T\widetilde{D_i}^T\left[\left.\frac{\partial E_{u|y}\Delta_iA_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right.\right.$$

$$
\begin{aligned}
&\quad -E^{(0)}\left\{\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\,\}\right]\widetilde{D_i}\,\xi\Bigg] \\
&\quad +\sum_{i=1}^{m}\left\{a^T\widetilde{D_i}^T\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}R_{ni}\right\} \\
&\quad +R_n^*(\mu^*) \\
&=: I_{n1}(\xi)+I_{n2}+I_{n3}(\xi).
\end{aligned}
$$

We will discuss the difference above step by step as follows:

It is obvious that (Y_i, U_i) are independent of one another for $i = 1, \ldots, m$ according to the models (2.1) and (2.2), then for $I_{n1}(\xi)$, by (A.2), (A.3) and (A.5), we have

$$
\begin{aligned}
E^{(0)}(I_{n1}(\xi))^2 &= E^{(0)}\Bigg[\sum_{i=1}^{p+N}a_k 1_k^T\sum_{i=1}^{m}\widetilde{D_i}^T\left\{\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right. \\
&\qquad\left.-E^{(0)}\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right\}\widetilde{D_i}\,\xi\Bigg]^2 \\
&\leqslant E^{(0)}\sum_{i=1}^{p+N}\Bigg[\sum_{i=1}^{m}1_k^T\widetilde{D_i}^T\left\{\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right. \\
&\qquad\left.-E^{(0)}\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right\}\widetilde{D_i}\,\xi\Bigg]^2 \\
&\leqslant \sum_{k,j}^{p+N}\sum_{i=1}^{m}E^{(0)}\Bigg[1_k^T\widetilde{D_i}^T\left\{\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right. \\
&\qquad\left.-E^{(0)}\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right\}\widetilde{D_i}1_j\Bigg]^2\|\xi\|^2 \\
&\leqslant \sum_{k,j}^{p+N}\sum_{i=1}^{m}(1_k^T\widetilde{D_i}^T\widetilde{D_i}1_k)E^{(0)}\|1_j^T\widetilde{D_i}^T\left(\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right. \\
&\qquad\left.-E^{(0)}\left.\frac{\partial E_{u|y}\Delta_i A_i^{-1/2}h_i(\mu_i)}{\partial\mu_i^T}\right|_{\mu_i=\mu_{0,i}}\Delta_{0,i}\right)\|^2\|\xi\|^2 \\
&\leqslant C\sup_i\sum_{k}^{p+N}(1_k^T\widetilde{D_i}^T\widetilde{D_i}1_k)\sum_{i}^{m}\sum_{k=1}^{p+N_k}(1_k^T\widetilde{D_i}^T\widetilde{D_i}1_k)\|\xi\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= C \sup_i trace(\widetilde{D_i}^T \widetilde{D_i})\, trace\left(\sum_{i=1}^m \widetilde{D_i}^T \widetilde{D_i}\right) \|\xi\|^2 \\
&\leqslant C k_n \sup_i trace(X_i^* K_n^{-1} X_i^{*T} + k_n \pi_i H_n^{-2} \pi_i^T) \|\xi\|^2 \\
&= O(\|\xi\|^2 k_n^2/n),
\end{aligned}
$$

where $1_k = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is a unit vector with 1 as its kth-element and 0 elsewhere and the finite constant C, independent of n, may vary from line to line. Thus, we have $E^{(0)}(I_{n1}(\xi))^2 = O(\|\xi\|^2 k_n^2/n)$. Consequently, for sufficiently large L,

$$
\sup_{\|\xi\| \leqslant L k_n^{1/2}, a^T a=1} |I_{n1}(\xi)| = O_p(n^{-1/2} k_n^{3/2}).
$$

For I_{n2}, we have

$$
\begin{aligned}
|I_{n2}| &\leqslant \left| \sum_{i=1}^m a^T \widetilde{D_i}^T \left[\left. \frac{\partial E_{u|y} \Delta_i A_i^{-1/2} h_i(\mu_i)}{\partial \mu_i^T} \right|_{\mu_i=\mu_{0,i}} \Delta_{0,i} \right.\right. \\
&\qquad \left.\left. - E^{(0)} \left. \frac{\partial E_{u|y} \Delta_i A_i^{-1/2} h_i(\mu_i)}{\partial \mu_i^T} \right|_{\mu_i=\mu_{0,i}} \Delta_{0,i} \right] R_{ni} \right| \\
&\qquad + \left| \sum_{i=1}^m a^T \widetilde{D_i}^T E^{(0)} \left. \frac{\partial E_{u|y} \Delta_i A_i^{-1/2} h_i(\mu_i)}{\partial \mu_i^T} \right|_{\mu_i=\mu_{0,i}} \Delta_{0,i} R_{ni} \right| \\
&= I_{n2}^{(1)} + I_{n2}^{(2)}.
\end{aligned}
$$

Similar to the proof of I_{n1}, we have $\sup_{a^T a=1} |I_{n2}^{(1)}| = O_p(k_n^{-r+1/2})$, $\sup_{a^T a=1} |I_{n2}^{(2)}| = O_p(k_n^{1/2})$.

For $I_{n3}(\xi)$, write $F_i^* = \Delta_i^T \frac{\partial^2 a^T \widetilde{D_i} E_{u|y} \Delta_i A_i^{-1/2} h_i}{\partial \mu_i^2} \Delta_i$ evaluated at $\mu_i = \mu_i^*$, by (A.3), (A.4) and (A.5), we have $\|F_i^*\| = O_p((k_n/n)^{1/2})$. And $I_{n3}(\xi)$ can also be expressed as

$$
\begin{aligned}
I_{n3}(\xi) &= \frac{1}{2} \sum_{i=1}^m \xi^T \widetilde{D_i}^T F_i^* \widetilde{D_i} \xi + \sum_{i=1}^m R_{ni}^T F_i^* \widetilde{D_i} \xi + \frac{1}{2} \sum_{i=1}^n R_{ni}^T F_i^* R_{ni} \\
&= I_{n3}^{(1)} + I_{n3}^{(2)} + I_{n3}^{(3)}.
\end{aligned}
$$

By the assumptions before, we have

$$
\begin{aligned}
&\sup_{\|\xi\| \leqslant L k_n^{1/2}, a^T a=1} |I_{n3}^{(1)}| = O_p(n^{-1/2} k_n^{5/2}), \\
&\sup_{\|\xi\| \leqslant L k_n^{1/2}, a^T a=1} |I_{n3}^{(2)}| = O_p(k_n^{3/2-r}), \\
&\sup_{\|\xi\| \leqslant L k_n^{1/2}, a^T a=1} |I_{n3}^{(3)}| = O_p(n^{1/2} k_n^{1/2-2r}). \qquad \text{(A.5)}
\end{aligned}
$$

So, $\sup_{\|\xi\| \leqslant L k_n^{1/2}, a^T a=1} |I_{n3}(\xi)| = O_p(k_n^{1/2})$.

Putting all the approximations together, we have

$$\sup_{\|\xi\| \leqslant Lk_n^{1/2}, a^T a=1} \|\Psi(\xi) - \Phi(\xi)\| = O_p(k_n^{1/2}), \tag{A.6}$$

and direct calculations give that

$$\|\widetilde{\xi}\| = O(k_n^{1/2}). \tag{A.7}$$

By (A.6) and (A.7), we have

$$\begin{aligned}\sup_{\|\xi\| \leqslant Lk_n^{1/2}} \|\Psi(\xi) - \xi\| &\leqslant \sup_{\|\xi\| \leqslant Lk_n^{1/2}} \|\Psi(\xi) - \Phi(\xi)\| + \|\widetilde{\xi}\| \\ &= LO_p(k_n^{1/2}) + O_p(k_n^{1/2}),\end{aligned} \tag{A.8}$$

which implies that

$$\sup_{\|\xi\| \leqslant Lk_n^{1/2}} \|\Psi(\xi) - \xi\| \leqslant Lk_n^{1/2},$$

in probability, for sufficiently large L.

Thus, Brouwer's fixed point theorem assures that the map $\xi \mapsto \xi - \Psi(\xi)$ has a fixed point $\hat{\xi}$, which is a zero of $\Psi(\xi)$, with $\|\hat{\xi}\| = O_p(k_n^{1/2})$, and arguments like those in He et al. [5] can now be used to prove (3.3).

Similar to the above arguments, we have

$$\sup_{\|\xi_1\| \leqslant L, \|\xi_2\| \leqslant k_n^{1/2}} \|\Psi_1(\xi_1, \xi_2) - \Phi_1(\xi_1, \xi_2)\| = o_p(1), \quad \|\hat{\xi}_1\| = O_p(1). \tag{A.9}$$

It follows from (A.9) that

$$\|\hat{\xi}_1 - \tilde{\xi}_1\| = o_p(1). \tag{A.10}$$

Therefore, to study the asymptotic normality of $\hat{\xi}_1 = K_n^{1/2}(\hat{\beta} - \beta_0)$, we shall show only the asymptotic normality of $\tilde{\xi}_1 = -E_{u|y}[\sum_{i=1}^m K_n^{-1/2} X_i^{*T} \Delta_{0,i} A_{0,i}^{-1/2} h_{0,i}(e_i)]$. By the central limit theorem, we can get the asymptotic normality of $\tilde{\xi}_1$. Thus Theorem 1 is proved. □

Proof of Theorem 2. First, we have

$$\begin{aligned} n^{-1}\hat{S}_n - S = \Bigg\{ & \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} X_{ij}^* E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) \\ & \times E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il}) X_{il}^{*T} \\ & - \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} X_{ij}^* E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) \\ & \times E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il}) X_{il}^{*T} \Bigg\} \end{aligned}$$

$$
\begin{aligned}
&+\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} X_{ij}^{*} E_{u|y}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})\right.\\
&\times E_{u|y}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})X_{il}^{*T}\\
&-\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} X_{ij}^{*} E_{u|y}^{(0)}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})\\
&\left.\times E_{u|y}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})X_{il}^{*T}\right\}\\
&+\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} X_{ij}^{*} E_{u|y}^{(0)}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})\right.\\
&\times E_{u|y}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})X_{il}^{*T}\\
&-\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} X_{ij}^{*} E_{u|y}^{(0)}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})\\
&\left.\times E_{u|y}^{(0)}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})X_{il}^{*T}\right\}\\
&+\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} X_{ij}^{*} E_{u|y}^{(0)}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})\right.\\
&\left.\times\, E_{u|y}^{(0)}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})X_{il}^{*T}-\frac{1}{n}S_n\right\}+\left\{\frac{1}{n}S_n-S\right\}\\
=:\ & C_{n1}+C_{n2}+C_{n3}+C_{n4}+C_{n5},
\end{aligned}
$$

where $r_{ij}=v_{ij}^{-1/2}(y_{ij}-\mu_{ij})$, $\hat{r}_{ij}$ denotes $r_{ij}(\theta_0)$ evaluated at $\hat{\theta}$. Note that $\mu_{ij}=g^{-1}(\eta_{0,ij}+\tilde{X}_{ij}^{T}\xi_1+\tilde{\pi}_{ij}^{T}\xi_2+R_{nij})$. By (A.1), for any d satisfying $\|d\|=1$, we have

$$
\begin{aligned}
&E^{(0)}\left|\sup_{\|d\|=1} d^{T}C_{n1}d\right|\\
&=E^{(0)}\left|\sup_{\|d\|=1}\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i} d^{T}X_{ij}^{*}X_{il}^{*T}d[E_{u|y}\dot{\mu}(\hat{r}_{ij})v^{-1/2}(\hat{r}_{ij})h(\hat{r}_{ij})\right.\\
&\times E_{u|y}\dot{\mu}(\hat{r}_{il})v^{-1/2}(\hat{r}_{il})h(\hat{r}_{il})\\
&\left.-E_{u|y}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})E_{u|y}\dot{\mu}(r_{il})v^{-1/2}(r_{il})h(r_{il})]\right|
\end{aligned}
$$

$$\leqslant E^{(0)} \sup_{\|d\|=1} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} \frac{|d^T X_{ij}^*|^2 + |X_{il}^{*T} d|^2}{2} |[E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij})$$

$$\times E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il}) - E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})]|$$

$$\leqslant \sup_{i,j \geqslant 1} n_i \sup_{\|d\|=1} d^T X_{ij}^* X_{ij}^{*T} d \sup_{i,j,l} E^{(0)} |[E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij})$$

$$\times E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il})$$

$$- E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})]|$$

$$\leqslant C \sup_{\|d\|=1} \frac{1}{n} d^T K_n d \sup_{i,j,l} E^{(0)} |[E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il})$$

$$- E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})]|.$$

Let $d_n = \sup_{ij}\{|\tilde{X}_{ij}^T \hat{\xi}_1 + \tilde{\pi}_{ij}^T \hat{\xi}_2 + R_{nij}|\}$, by $\|\hat{\xi}\| = O_p(k_n^{1/2})$, (A.5) and Lemma A.1, we have

$$d_n = O_p(n^{-3/10}) = o_p(n^{-1/5}). \tag{A.11}$$

For sufficiently large $C > 0$, $P(d_n \leqslant C n^{-1/5}) \to 1, n \to \infty$. By the continuity of bounded function $E_{u|y} \dot{\mu}_i(.) v^{-1/2}(.) h(.)$, we have

$$E^{(0)} \Bigg\{ I(d_n \leqslant C n^{-1/5}) \sup_{i,j,l} |E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il})$$

$$- E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})| \Bigg\}$$

$$= E^{(0)} \Bigg[I(d_n \leqslant C n^{-1/5}) \Bigg\{ \sup_{i,j,l} |E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il})$$

$$- E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})$$

$$+ E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})$$

$$- E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij}) E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})| \Bigg\} \Bigg]$$

$$\leqslant C \sup_{i,j,l} [E^{(0)} \{E_{u|y} \dot{\mu}(\hat{r}_{il}) v^{-1/2}(\hat{r}_{il}) h(\hat{r}_{il}) - E_{u|y} \dot{\mu}(r_{il}) v^{-1/2}(r_{il}) h(r_{il})\}^2]^{1/2}$$

$$+ C \sup_{i,j,l} [E^{(0)} \{E_{u|y} \dot{\mu}(\hat{r}_{ij}) v^{-1/2}(\hat{r}_{ij}) h(\hat{r}_{ij}) - E_{u|y} \dot{\mu}(r_{ij}) v^{-1/2}(r_{ij}) h(r_{ij})\}^2]^{1/2} \longrightarrow 0.$$

Thus,

$$\|C_{n1}\| = o_p(1). \tag{A.12}$$

Similar to the proof of (A.11), it is easy to verify that

$$\|C_{n2}\| = o_p(1) \quad \text{and} \quad \|C_{n3}\| = o_p(1). \tag{A.13}$$

By the independency of (Y_i, U_i), we get

$$Var\left(\sup_{\|d\|=1} d^T C_{n4} d\right) \leqslant \frac{C}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sup_{\|d\|=1}|d^T X_{ij}^*|^4 E^{(0)}|E_{u|y}\dot{\mu}(r_{ij})v^{-1/2}(r_{ij})h(r_{ij})|^4 \longrightarrow 0. \tag{A.14}$$

Since $E^{(0)}(C_{n4}) = 0$, we have

$$\|C_{n4}\| = o_p(1). \tag{A.15}$$

By (A.12), (A.13), (A.15) and (A.5),

$$n^{-1}\hat{S}_n - S = o_p(1). \tag{A.16}$$

Similar to (A.16) and by (A.5), we have

$$n^{-1}\hat{K}_n - K = o_p(1). \tag{A.17}$$

Thus, by (A.16) and (A.17), Theorem 2 is proved. □

References

[1] J.G. Booth, J.P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, J. Roy. Statist. Soc. Ser. B 61 (1999) 265–285.

[3] W. Härdle, E. Mammen, M. Müller, Testing parametric versus semiparametric modelling in generalized linear models, J. Amer. Statist. Assoc. 93 (1998) 1461–1474.

[4] X. He, W.K. Fung, Z.Y. Zhu, Robust estimation in generalized partial linear models for clustered data, J. Amer. Statist. Assoc. 100 (2005) 1176–1184.

[5] X. He, Z.Y. Zhu, W.K. Fung, Estimation in a semiparametric model for longitudinal data with unspecified dependence structure, Biometrika 89 (2002) 579–590.

[6] P.J. Huber, Robust Statistics, Wiley, New York, 1981.

[7] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.

[8] X. Lin, R.J. Carroll, Semiparametric regression for clustered data, Biometrika 88 (2001) 1179–1185.

[9] X. Lin, R.J. Carroll, Semiparametric regression for clustered data using generalized estimating equations, J. Amer. Statist. Assoc. 96 (2001) 1045–1056.

[10] P. McCullagh, J. Nelder, Generalized Linear Models, second ed., Chapman & Hall, London, 1989.

[11] C.E. McCulloch, Maximum likelihood algorithms for generalized linear mixed models, J. Amer. Statist. Assoc. 92 (1997) 162–170.

[12] M. Müller, Estimation and testing in generalized partial linear models—a comparative study, Statist. Comput. 11 (2001) 299–309.

[13] J.S. Preisser, B.F. Qaqish, Robust regression for clustered data with application to binary responses, Biometrics 55 (1999) 574–579.

[14] J. Rice, Convergence rates for patially splined models, Statist. Probab. Lett. 4 (1986) 203–208.

[15] P.J. Rousseeuw, B.C. van Zomerem, Unmasking multivariate outliers and leverage points, J. Amer. Statist. Assoc. 85 (1990) 633–639.

[16] L.L. Schumaker, Spline Functions, Wiley, New York, 1981.

[18] T.A. Severini, J.G. Staniswalis, Quasi-likelihood estimation in semiparametric models, J. Amer. Statist. Assoc. 89 (1994) 501–511.

[19] S.K. Sinha, Robust analysis of generalized linear mixed models, J. Amer. Statist. Assoc. 99 (2004) 451–460.

[20] R. Stiratelli, N. Laird, J. Ware, Random effects models for serial observations with binary responses, Biometrics 40 (1984) 961–971.

[21] C. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.

[22] M.A. Tanner, Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, second ed., Springer, New York, 1993.
[23] Y.G. Wang, X. Lin, M. Zhu, Robust estimation functions and bias correction for longitudinal data analysis, Biometrics 61 (2005) 684–691.
[24] S.L. Zeger, M.R. Karim, Generalized linear models with random effects; A Gibbs sampling approach, J. Amer. Statist. Assoc. 86 (1991) 79–86.
[25] S.L. Zeger, K.Y. Liang, P.S. Albert, Models for longitudinal data: a generalized estimating equation approach, Biometrics 44 (1988) 1049–1060.